Н. И. Киселёв

Альтернативные методы оценки главных компонент

Критерии минимаксного типа рассматриваются как альтернатива методу наименьших квадратов в определении главных компонент. Оценка коэффициентов формулируется как задача линейного программирования. Предложенный подход экспериментально проверяется на известных тестовых статистических массивах. На этих данных получены результаты, не уступающие оценкам классического метода наименьших квадратов, а в некоторых задачах и превосходящие их.

Ключевые слова: главные компоненты, минимаксные критерии, линейное программирование.

1. Введение

В методе главных компонент не предполагается в явном виде, что исходные данные являются наблюдениями случайной величины с многомерным нормальным распределением, но использование метода наименьших квадратов, ковариационной матрицы, геометрических представлений типа «эллипсоид рассеяния» и т. п. выдают подсознательное предположение статистиков о модели нормального распределения. Другие модели образования исходных данных, рассматриваемые в контексте выбора метода оценивания главных компонент, в литературе не найдены. Однако в эконометрике нормальное распределение данных не является доминантой, как это было ранее в других приложениях статистических методов. Более того, в относительно простых задачах, например, при оценивании центра случайной величины по выборке, помимо выборочного среднего используется множество других оценок, в частности, медиана, середина размаха и т. д., каждая из которых имеет оптимальные свойства в соответствующих распределениях случайной величины.

Для развития множественности подходов к решению важных задач математической статистики представляется целесообразным построить альтернативу классическому методу главных компонент. В частности, вместо среднеквадратичного отклонения как меры изменчивости переменных можно использовать максимальный размах, вместо критерия наименьших квадратов — минимаксный критерий отклонения, а вместо многомерного эллипсоида как геометрического образа формы данных использовать прямоугольный параллелепипед. Очевидно, что в зависимости от модели данных прикладной задачи тот или иной метод изучения геометрии многомерных наблюдений будет иметь свои преимущества.

Известно, что при использовании классического метода определения главной компоненты одновременно оптимизируются два критерия: критерий минимума суммы квадратов отклонений наблюдений от этой компоненты и критерий максимума суммы квадратов значений их проекций на компоненту (Айвазян и др., 1989). Однако при выборе максимального размаха как меры отклонения наблюдений в общем случае будут получаться различные главные компоненты при использовании критерия минимаксного размаха или максимальной проекции.

В этой связи, оба указанных критерия далее будут рассмотрены и применены к известным тестовым данным (выборки данных по видам цветка ириса и по макроэкономическим показателям России за 1995–2008 гг.). Для изучения их свойств приводится сравнение полученных оценок с результатами классического метода главных компонент.

Помимо основной задачи определения главных компонент, минимаксный подход дает возможность решить задачу локализации многомерных данных, а именно задачу построения параллелепипеда, который содержит все наблюдения, причем его форма и угловая ориентация в пространстве определяют геометрию исходных данных. Более того, относительный объем параллелепипеда можно использовать как меру общей линейной связи наблюдений.

2. Определение главных компонент из условия максимального размаха (1-ый метод)

Пусть X — матрица числовых данных размером $n \times m$; m — число показателей, регистрируемых в каждом наблюдении, n — число наблюдений, $n \geq m + 1$, матрица X имеет ранг m. Наблюдения $x^{(1)}, x^{(2)}, \ldots, x^{(n)}$ образуют $n(n-1)/2$ различных пар $(x^{(k)}-x^{(l)})$, $k > l$. В качестве меры расстояния (размаха) между наблюдениями пары будем использовать обычную евклидову метрику $\| x^{(k)} - x^{(l)} \|$.

Определим первую главную компоненту как направление (вектор $\hat{c}_1$), на котором достигается максимальная величина проекции среди всех пар наблюдений, т. е.

$$\hat{c}_1 = \arg\max_{c} \max_{k,l} \left| c(x^{(k)} - x^{(l)}) \right|, \quad \| c \| = 1, \tag{1}$$

где $c(x^{(k)} - x^{(l)})$ — скалярное произведение, что при условии нормировки $\| c \| = 1$ является величиной проекции вектора $(x^{(k)} - x^{(l)})$ на направление c.

В такой постановке (1) направление первой главной компоненты очевидно, оно будет совпадать с вектором, соединяющим два наблюдения, расстояние между которыми максимально. Тогда наибольшая величина проекции на главную компоненту будет равна максимальному расстоянию между парами в выборке и, очевидно, что не существует другого направления с большей величиной проекции. Если пар наблюдений с максимальным расстоянием несколько, то решение не единственно (этот случай здесь рассматриваться не будет). Другими словами, определение первой главной компоненты сводится к нахождению пары наблюдений с максимальным расстоянием. Вектор, соединяющий эту пару, указывает направление главной компоненты, а его длина является нагрузкой на нее (аналог собственному значению в классическом случае).

В общем случае i-ая главная компонента *вычисляется по следующему алгоритму.* Пусть $\hat{c}_1, \hat{c}_2, \ldots, \hat{c}_{i-1}$ — нормированные вектора вычисленных ранее главных компонент. Проектируем все наблюдения на пространство, образованное указанными векторами. В результате получаем вектора проекций $x^{(1)}_{pr\|}, x^{2}_{pr\|}, \ldots, x^{(n)}_{pr\|}$ наблюдений. Из разложения

$$x^{(j)} = x^{(j)}_{pr\|} + x^{(j)}_{pr\perp}, \quad j = 1, \ldots, n$$

получаем последовательность проекций наблюдений $x^{(1)}_{pr\perp}, x^{2}_{pr\perp}, \ldots, x^{(n)}_{pr\perp}$ на пространство, ортогональное к ранее найденным главным компонентам. Далее следуют действия, ана-

логичные вычислению первой компоненты: полученные проекции образуют, как и выше, набор из $n(n-1)/2$ различных пар, и среди них находим пару с максимальным расстоянием. Вектор, соединяющий эту пару, определяет направление, а его величина — нагрузку i-ой главной компоненты.

Заметим, что проекционная матрица P_{i-1} для получения последовательности $x_{pr\|}^{(1)}, x_{pr\|}^{(2)}, \ldots, x_{pr\|}^{(n)}$ на i-ой итерации в случае ортонормированных векторов главных компонент имеет простой вид

$$P_{i-1} = C_{i-1} C_{i-1}^T,$$

где C_{i-1} — матрица, составленная из векторов $\hat{c}_1, \hat{c}_2, \ldots, \hat{c}_{i-1}$ главных компонент размером $m \times (i-1)$.

Для подтверждения дееспособности данного подхода рассмотрим его применение к классическим данным цветков ириса (Fisher, 1936), которые зачастую используются для тестирования методов анализа многомерных данных, в частности, при распознавании образов. Данные заимствованы из открытого источника (Asuncion, Newman, 2007) репозитория UCI тестовых статистических массивов, организованного в университете г. Ирвайн (Калифорния, США). Содержательно это выборка из 150 наблюдений цветков ириса, для каждого измерены 4 классификационных ботанических признака. Цветки ириса принадлежат к трем видам, которые в выборке представлены подвыборками по 50 наблюдений каждого вида. Известно, что один вид линейно отделяется от двух других, которые, однако, сами не имеют линейного дискриминатора.

Эксперимент состоит в проверке — повторит ли изложенный выше подход известные результаты по разделению видов цветка ириса в координатах его первых двух компонент. Исходные ботанические данные предварительно масштабировались путем деления измеренных значений признака на его максимальный размах. Результаты метода максимального размаха приведены на рис. 1. Для сравнения на рис. 2 даны представления наблюдений в первых двух

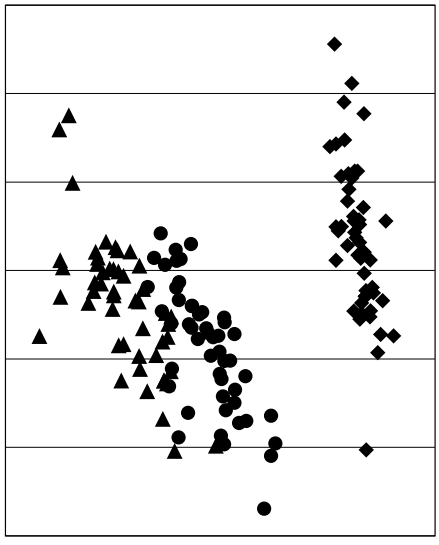

Рис. 1. Представление данных методом максимального размаха

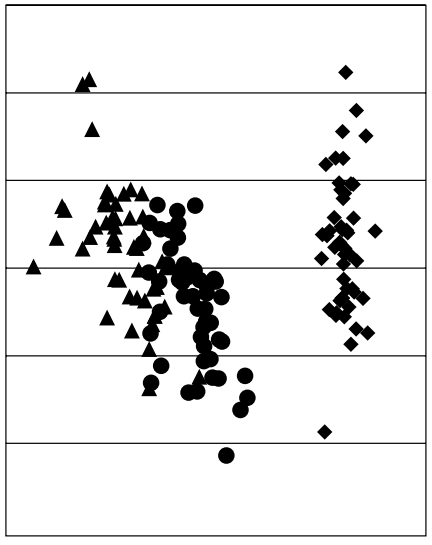

Рис. 2. Представление данных классическим методом

компонентах по классическому методу главных компонент. Результаты по классическому методу рассчитаны автором, аналогичные результаты приведены ранее в (Зиновьев, 2000).

Обозначения на рисунках наблюдений: ▲ — Iris-versicolor, ● — Iris-virginica, ◆ — Iris-setosa.

Видно, что качественно рисунки близки, однако на рис. 1 наблюдения различных классов «разнесены» сильнее, чем на рис. 2. Другими словами, представление данных в координатах первых двух компонент, полученных методом максимального размаха, имеет несколько более четко выраженную классификационную структуру.

Подтверждением близости результатов обоих методов в данном примере является табл. 1, где приведены значения косинусов углов между векторами главных компонент и каждой осью координат (значения для классического метода даны вторым числом через «слэш»). Курсивом обозначены ячейки с большими значениями косинусов. Как видно, *их значения весьма близки для обоих методов.*

Таблица 1. Значение косинусов углов векторов главных компонент, полученных методом максимального размаха и классическим методом

№ вектора	x_1	x_2	x_3	x_4
1	*0.57 / 0.52*	−0.10 / −0.26	*0.59 / 0.58*	*0.55 / 0.57*
2	0.36 / 0.37	*0.92 / 0.92*	−0.13 / 0.02	−0.06 / 0.06
3	*0.70 / 0.72*	−0.35 / −0.24	−0.19 / −0.14	*−0.59 / −0.63*
4	0.20 / 0.26	−0.15 / −0.12	*−0.77 / −0.80*	*0.58 / 0.52*

Отличие результатов этих методов имеют место лишь в нагрузках на каждую компоненту. Для рассматриваемого метода нагрузки для всех четырех компонент составляют: 48%, 29%, 15%, 8% (рассчитывались как отношение размаха по данной компоненте к суммарному размаху по всем компонентам). Для классического метода, соответственно, получаем: 73%, 22.5%, 4%, 0.5% (процентные значения собственных значений ковариационной матрицы). Казалось бы, классический метод предпочтительней в смысле распределения нагрузки по главным компонентам. Однако собственные значения — это квадратичная функция от исходных данных, тогда как размах зависит от них линейно. Если извлечь корень из собственных значений и вновь вычислить процентное соотношение, то получим: 53%, 30%, 12.5%, 4.5%. Видно, что снова для обоих методов получаются близкие результаты.

3. Определение главных компонент из условия минимума максимального размаха (2-ой метод)

Используя обозначения, введенные в предыдущем разделе, запишем условие определения направления $\hat{c}_1$ первой компоненты (по величине изменчивости это будет последняя компонента), для которой имеет место минимум максимальной величины проекции на нее среди всех значений проекций пар наблюдений

$$\hat{c}_1 = \arg\min_{c} \max_{k,l} \left| c(x^{(k)} - x^{(l)}) \right|, \quad \| c \| = 1, \tag{2}$$

где, как и ранее, $c(x^{(k)} - x^{(l)})$ — скалярное произведение, которое при условии нормировки $\| c \| = 1$ дает значение проекции вектора $(x^{(k)} - x^{(l)})$ на направление c.

В постановке (2) вектор $\hat{c}_1$ означает направление, на котором имеет место минимальная изменчивость (размах), тогда как в классическом случае первая главная компонента является направлением максимальной изменчивости. Однако в общем случае при изучении геометрии данных нет формальных предпочтений, в каком порядке строить главные компоненты: от максимальной изменчивости к минимальной или в обратную сторону. На наш взгляд, этот порядок определяется содержанием прикладной задачи. В данном случае начало построения главных компонент из условия минимальной изменчивости следует из постановки (2).

Решение задачи оптимизации (2) при, казалось бы, близком к задаче (1) виде, существенно сложнее. Рассмотрим несколько более простую задачу оптимизации (3), решение которой даст нам искомый вектор $\hat{c}_1$ из (2)

$$\hat{c}_1 = \arg \min_{c, c_0} \max_i \left| cx^{(i)} + c_0 \right|, \ \| c \| = 1, \ i = 1, \ldots, n, \tag{3}$$

где c_0 — свободный член уравнения гиперплоскости $cx + c_0 = 0$. Задача (3) означает поиск таких значений коэффициентов плоскости, при которых максимальное значение отклонения наблюдений от плоскости будет минимальным. Другими словами — это определение минимаксной плоскости, где отклонением точки служит ее евклидово расстояние до плоскости. Несложно показать, что направляющий вектор оптимальной плоскости из (3) будет являться решением задачи (2).

В самом деле, пусть имеется решение задачи (3) — минимаксная плоскость с максимальным отклонением b и направляющим вектором c. Тогда проекции точек на этот вектор лежат на отрезке $[b, -b]$, если предположить, что 0 является пересечением вектора и плоскости. Следовательно, максимальная величина проекций пар точек на вектор равна $2b$. Предположим, что решение задачи (2) дает другой направляющий вектор с величиной максимальной проекции пар наблюдений $2\tilde{b}$, меньшей, чем $2b$. Тогда плоскость, ортогональная этому вектору и проходящая через середину отрезка $2\tilde{b}$, будет иметь максимальное отклонение наблюдений от плоскости $\tilde{b}$ меньше, чем b, что противоречит условию задачи (3).

Для решения задачи (3) будем использовать аппарат линейного программирования (ЛП) (Гольштейн, 1971). Введем искусственные неотрицательные переменные (невязки) b_i^+ и b_i^- для каждого наблюдения и переменную b максимального отклонения по всем наблюдениям. Тогда эту задачу можно записать в терминах ЛП в канонической форме, но с одним квадратичным ограничением, связанным с условием нормировки:

$$b \to \min_{c, c_0} \tag{4}$$

$$Xc + c_0 + be - b^- = 0, \ b^- \geq 0, \tag{5}$$

$$Xc + c_0 - be + b^+ = 0, \quad b^+ \geq 0, \quad b \geq 0, \tag{6}$$

$$\| c \| = 1, \tag{7}$$

где e — единичный вектор, b^+ и b^- — вектора невязок размерности n. Как видно из (4)–(7), задача оптимизации имеет $2n + 1$ нетривиальных ограничений и $2n + m + 2$ неизвестных переменных.

Поясним содержание ограничений (5)–(6). В ограничениях (5) (их n по числу наблюдений) к значению $cx^{(i)} + c_0,\ i = 1,\ldots,n$ каждого наблюдения прибавляется такая величина b, что их суммарное значение больше или равно нулю. Это следует из условия $b^- \geq 0$ и геометрически означает, что все наблюдения будут лежать выше или на плоскости $cx + c_0 = 0$. Соответственно, в ограничениях (6) величина b вычитается, и тогда из условия $b^+ \geq 0$ следует, что все наблюдения будут лежать ниже или на плоскости. Цель решения задачи (4)–(7) состоит в нахождении таких параметров плоскости $cx + c_0 = 0$, чтобы при данных наблюдениях (матрица X) величина b была наименьшая. В этом случае, очевидно, что b является искомым наименьшим максимальным отклонением наблюдений от плоскости $cx + c_0 = 0$.

В результате решения (4)–(7) получаем искомый направляющий вектор $\hat{c}_1$ и значение максимального отклонения $\hat{b}_1$ (нижний индекс 1 в этих переменных означает, что они относятся к первой компоненте). Так как одновременно находится и оценка свободного члена $\hat{c}_{01}$, то первая минимаксная плоскость полностью определена. Помимо этого, с помощью величины $\hat{b}_1$ задаются две параллельные ей плоскости, отстоящие от минимаксной на эту величину и содержащие между собой все наблюдения (в том числе, лежащие на самих граничных плоскостях).

Для i-ой главной компоненты повторяется решение задачи (4)–(7), но с дополнительными условиями ортогональности искомого оптимального решения задачи найденным на предшествующих шагах направляющим векторам.

$$\widehat{C}_{i-1} c = 0. \tag{8}$$

Строки матрицы $\widehat{C}_{i-1}$ состоят из оценок коэффициентов минимаксной плоскости, полученных на предыдущих шагах (без оценки свободного члена), т. е. если таких шагов $i - 1$, то матрица $\widehat{C}_{i-1}$ имеет размерность $(i-1) \times m$, и ее строки попарно ортогональны.

В результате решения (4)–(8) получаем вектор $\hat{c}_i$ и i-ую пару плоскостей с расстоянием $\hat{b}_i$ между ними, вновь содержащих между собой все наблюдения и ортогональных ранее полученным парам плоскостей.

Пересечением m пар взаимно ортогональных плоскостей, полученных при определении всех главных компонент, является прямоугольный параллелепипед, содержащий все наблюдения. Этот параллелепипед *обеспечивает простую геометрическую интерпретацию главным компонентам*. Он однозначно определяется набором m различных векторов ребер, исходящих из одной вершины, которые можно ассоциировать с главными компонентами. Следуя общепринятому порядку перечисления компонент (где первая компонента указывает максимальную изменчивость), имеем, соответственно: *максимальное по длине ребро параллелепипеда задает направление и величину (длина ребра) первой главной компоненты, второе по длине — вторую и т. д.*

Возвращаясь к технике решения задачи (4)–(8), заметим, что вследствие условия нормировки $\| c \|^2 = 1$ использовать непосредственно симплекс метод решения задачи ЛП не представляется возможным. В этом случае задача решается итеративным процессом, суть которого состоит в замене на k-ой итерации ограничения $\| c \| = 1$ на линейное условие

$$c^T \hat{c}_i^{(k-1)} = 1, \tag{9}$$

где $\hat{c}_i^{(k-1)}$ — решение задачи определения i-ой компоненты на $(k-1)$-ой итерации. В качестве значения $\hat{c}_i^{(0)}$ для первой итерации предпочтительней выбрать решение, полученное первым методом для соответствующей компоненты. В этом случае решение задачи (4)–(6), (8)–(9) находится за одну итерацию, а вторая необходима для срабатывания правила остановки процесса, которое в изложенном ниже эксперименте имеет вид:

$$\left|\hat{b}_i^{(k)} - \hat{b}_i^{(k-1)}\right| / \hat{b}_i^{(k)} \le \alpha,$$

где $\hat{b}_i^{(k-1)}$ — значение максимального отклонения на $(k-1)$-ой итерации вычисления i-ой компоненты, α — малая величина (в эксперименте ниже 0.01). В случае выбора равноугольного начального условия $\hat{c}_i^{(0)} = 1/\sqrt{m}$ количество итераций увеличивается незначительно (на одну-две).

Как видно из (4)–(6), (8)–(9), задача ЛП при вычислении i-ой компоненты имеет $2n+i-1$ нетривиальных ограничений и $2n+m+1$ неизвестных переменных, что при значительных объемах наблюдений может приводить к большим объемам вычислений. С этой точки зрения задача ЛП, двойственная к задаче (4)–(6), (8)–(9), требует меньших объемов вычислений и, более того, двойственные оценки содержат, как будет показано ниже, индикаторную информацию о самих точках.

Сформулируем задачу ЛП, двойственную к задаче оптимизации (4)–(6), (8)–(9) в случае вычисления i-ой компоненты на k-ой итерации:

$$\gamma \to \min_{\lambda,\mu,\nu}, \tag{10}$$

$$e^T(\lambda - \mu) = 0, \tag{11}$$

$$X^T(\lambda - \mu) + \widehat{C}_{i-1}^T \nu + \gamma \hat{c}_i^{(k-1)} = 0, \tag{12}$$

$$e^t(\lambda + \mu) \le 1, \tag{13}$$

$$\lambda \ge 0, \mu \ge 0,$$

где λ и μ — вектора двойственных переменных размерности n, относящиеся, соответственно, к ограничениям (5) и (6) прямой задачи, вектор ν имеет размерность $i-1$, его компоненты не ограничены в знаке и являются двойственными оценками для ограничений (8), переменная γ относится к ограничению (9). Отметим, что число ограничений (10)–(13) всего лишь $m+2$.

Минимаксная плоскость и пара граничных плоскостей, которые она индуцирует, имеют в многомерном случае свойства, известные по одномерной ситуации: все наблюдения находятся между минимальным и максимальным значениями, выборочная оценка центра — средняя точка находится на равном расстоянии от этих значений и устойчива к колебаниям внутренних точек выборки. В многомерном случае роль минимальных и максимальных значений выполняют пары граничных плоскостей и опорные точки (т. е. наблюдения, лежащие на плоскостях), число которых для каждой пары не менее $m+1$ (это справедливо в случае отсутствия условий

(8) ортогональности; каждое условие ортогональности уменьшает число $m+1$ на единицу). Остальные точки — внутренние, и если при их колебаниях они не выходят из области, ограниченной парой параллельных плоскостей, то минимаксная плоскость остается прежней, т. е. в указанном смысле эта плоскость устойчива к внутренним точкам. Эти и другие свойства минимаксной регрессии более подробно рассматривались в (Киселев, 1985).

Двойственные оценки λ_i и μ_i относятся к *i*-му наблюдению и являются его важной характеристикой. Если в оптимальном решении (10)–(13) при вычислении *j*-ой компоненты имеем $\lambda_i = \mu_i = 0$, то *i*-ое наблюдение лежит внутри пары минимаксных плоскостей этой компоненты. Если $\lambda_i \neq 0$, то $\mu_i = 0$, из этого следует, что *i*-ое наблюдение является опорным и лежит на одной из плоскостей пары. Если $\lambda_i = 0$ и $\mu_i \neq 0$, то точка лежит на другой плоскости этой пары. Область значений λ_i и μ_i является отрезком [0, 0.5], при этом (см. (11) и (13)) сумма по всем λ_i равна 0.5 и равна сумме по всем μ_i. Заметим, что в задачах оценки макрохарактеристики линейной связи показателей (см. раздел 4) двойственные оценки наблюдений можно интерпретировать как меру их аномальности, связанной с отклонением наблюдений от линейной связи.

Для проверки применимости минимаксного подхода к определению главных компонент использовались статистические данные, взятые из открытого источника http://data.cemi.rssi.ru/GRAF/InpDat.php «Эконометрическая модель экономики России» (В. Макаров, С. Айвазян и др.) на сайте ЦЭМИ РАН. Данные представляют поквартальные наблюдения, начиная с четвертого квартала 1995 года по 2008 год включительно, следующих четырех показателей:

x_1 — значение валового внутреннего продукта (ВВП);

x_2 — величина инвестиций с лагом в 4 квартала;

x_3 — квартальное приращение курса доллара;

x_4 — значение ВВП с лагом в один квартал.

Таким образом, фактические данные представляют 53 точки в четырехмерном пространстве: $n = 53$ и $m = 4$.

Эмпирические распределения рассматриваемых показателей приведены на рис. 3. Следует отметить, что визуально они весьма отличаются от нормального. В этом случае применение классического метода главных компонент с использованием свойств нормального распределения (средние значения, ковариационная матрица и т. д.) не является обоснованным.

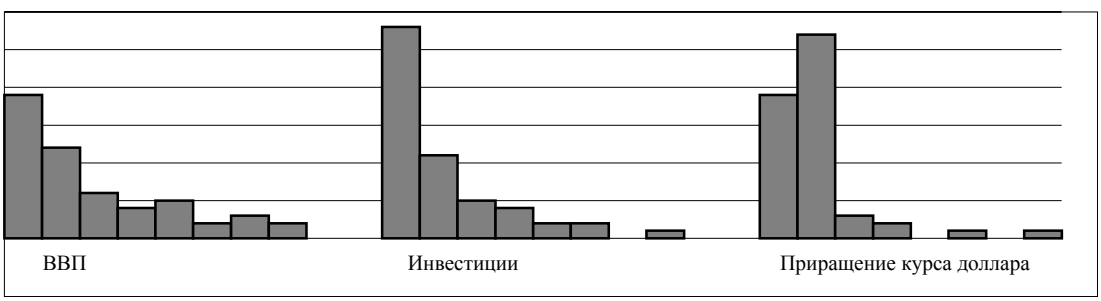

Рис. 3. Эмпирические плотности распределений показателей

В результате применения изложенной выше процедуры определения главных компонент путем многократного решения задачи (10)–(13) получен параллелепипед, косинусы углов

между ребрами которого (направление компонент) и осями координат приведены в табл. 2. Длины ребер параллелепипеда в процентном отношении равны: первое ребро (первая главная компонента) по длине составляет 61% от общей суммы всех четырех ребер, второе (вторая компонента) — 30% и два последних — по 5% и 3%. Заметим, что исходные измерения показателей предварительно масштабировались путем деления значений каждого показателя на его размах и центрировались вычитанием его минимального значения.

Для сравнения по этим данным вычислена их ковариационная матрица и найдены ее собственные значения и вектора, т. е. главные компоненты определены классическим методом. В процентном отношении эти собственные значения следующие: первое составляет 88.4%, второе — 11%, третье — 0.5% и четвертое — 0.1%. Как и в первом эксперименте, переходим от квадратичной характеристики изменчивости к линейной (т. е. извлекаем корень из собственных значений и пересчитываем процентные соотношения). В результате получаем: первая компонента содержит 68.5%, вторая — 24%, третья — 5% и четвертая — 2.5%, что вновь близко к минимаксному методу.

Также главные компоненты вычислялись первым методом из условия максимального размаха. Представим в виде триады процентные отношения для всех трех методов. Первая тройка чисел — это процент нагрузки первой главной компоненты для первого, второго и классического методов соответственно, вторая тройка чисел — это процент второй компоненты и т. д.

57% — 61% — 68.5%, 34% — 30% — 24%, 6% — 6% — 5%, 3% — 3% — 2.5%.

Как видно из приведенных значений, в классическом методе суммарная изменчивость на направлениях первого и второго собственного векторов практически равна изменчивости по первым двум «длинным» ребрам параллелепипедов рассматриваемых методов.

Матрица нормированных векторов ребер (косинусы углов между векторами и осями координат) приведена в табл. 2, где первое и второе число в ячейке относится к первому и второму минимаксному методу, а третье число — косинусы собственных векторов для классического случая. Курсивом выделены ячейки, где имеют место минимальные углы с соответствующими осями. Отметим, что в данном примере выделенные ячейки совпадают для всех методов на первых двух компонентах.

Таблица 2. Косинусы углов между главными компонентами и осями координат для трех методов

№ вектора	x_1 ВВП	x_2 Инвестиции	x_3 Изменение курса доллара	x_4 ВВП с лагом
1	*0.66 / 0.70 / 0.71*	0.10 / 0.12 / 0.11	−0.47 / −0.33 / −0.10	*0.58 / 0.63 / 0.68*
2	0.24 / 0.22 / 0.06	0.13 / 0.14 / 0.05	*0.87 / 0.94 / 0.99*	0.41 / 0.22 / 0.07
3	0.66 / 0.62 / 0.17	0.26 / 0.24 / 0.98	0.11 / 0.0 / 0.03	−0.70 / −0.74 / −0.01
4	−0.29 / −0.28 / −0.67	0.95 / 0.95 / −0.12	−0.10 / −0.10 / 0.01	0.07 / 0.08 / 0.72

Из таблицы косинусов следует, что значения первого самого длинного ребра определяется, в основном, первым и четвертым показателями (ВВП и ВВП с лагом в один квартал),

причем вклад каждого из них примерно одинаков. Второе ребро имеет весьма малый угол с показателем «приращение курса доллара», т. е. в основном связано с этим показателем. Нагрузки на третье и четвертое ребро для минимаксных методов и классического метода расходятся и, если пытаться их интерпретировать, получаются разные версии.

Сравнение первого и второго методов показывает, что они хорошо согласуются на всех компонентах. Сравнение этих методов с классическим показывает, в целом, хорошую согласованность на первых двух векторах. На двух последних компонентах, которые учитывают малую долю изменчивости, согласования между минимаксными и классическим методом не наблюдается.

В целом, эксперимент на использованных реальных данных демонстрирует, на наш взгляд, разумные результаты, во многом хорошо согласованные с расчетами классическим методом наименьших квадратов. Для изучения эффективности минимаксного подхода требуются, естественно, дополнительные теоретические исследования и эксперименты, которые определят области предпочтительного применения метода.

Заметим, однако, что минимаксный подход следует рассматривать не только как альтернативу классическим главным компонентам, представляется, что он дает дополнительную полезную информацию. В частности, как уже отмечалось, локализация многомерных данных в простом геометрическом образе (параллелепипеде), на гранях которого находится часть наблюдений, позволяет получить ряд содержательно интересных результатов.

4. Относительный объем параллелепипеда в определении меры общей линейной связи показателей

Помимо параллелепипеда (обозначим его P_1), построенного выше как пересечение m пар минимаксных плоскостей, рассмотрим другой прямоугольный параллелепипед P_0, ребра которого равны, соответственно, размахам по каждому показателю, а грани перпендикулярны координатным осям. Параллелепипед P_0 является также пересечением m пар параллельных плоскостей, каждая из которых определяется по одному показателю без учета их взаимной зависимости, тогда как при построении P_1 учитывается по существу линейная связь показателей. В этом случае можно ввести *новую макрохарактеристику линейной связи* ρ всей совокупности рассматриваемых показателей как величину относительного объема параллелепипеда P_1 в P_0.

$$\rho = 1 - (V_{P_1} / V_{P_0})^{1/m},$$

где V_{P_1} и V_{P_0} объемы, соответственно, параллелепипедов P_1 и P_0. Извлечение корня степени m (размерность пространства показателей) из отношения объемов элиминирует влияние размерности. Значение ρ находится на отрезке [0, 1], где значение 1 указывает на наличие, как минимум, одной строгой линейной зависимости между показателями, а значение 0 соответствует их независимости

Как известно, в общем случае объем косоугольного параллелепипеда равен модулю детерминанта матрицы, состоящей из векторов его ребер, исходящих из одной вершины. В нашем случае, при определении параллелепипеда как пересечения m пар плоскостей, удобно его объем вычислять по формуле

$$V_p = 2^m b_1 \ldots\ldots b_m \Big/ \det[c_{ij}]_{i,j}^m \,, \qquad (14)$$

где, как и выше, b_i — максимальное отклонение наблюдений от i-ой минимаксной плоскости и матрица $[c_{ij}]_{i,j}^m$ состоит из направляющих векторов минимаксных плоскостей. Для прямоугольных параллелепипедов знаменатель в (14) равен 1. Формулы для объемов параллелепипеда (14) и эллипсоида (15) (см. ниже) приведены на сайте http://www.pmpu.ru/vf4/dets/geometry/ автора А. Ю. Утешева.

В случае эксперимента на макроэкономических данных, отношение собственно объемов P_1 / P_0 равно 0.11 для первого метода и 0.07 для второго. Эти значения, казалось бы, указывают на высокую степень линейной связи. Однако вычисление коэффициента линейной связи ρ, который учитывает влияние размерности, приводит, соответственно, к значениям 0.42 и 0.49, что указывает уже не на высокую, но заметную линейную связь рассматриваемых показателей в исходных данных.

Представляет интерес сравнение объема полученного параллелепипеда с объемом минимально возможного эллипсоида, содержащего все наблюдения. Методика сравнения следующая. По данным эксперимента вычисляется выборочная ковариационная матрица и находится обратная ей матрица Q. Искомый эллипсоид с центром в выборочной оценке средних $\hat{\mu}$ определяется как

$$(x - \hat{\mu})^T Q (x - \hat{\mu}) = D^2,$$

здесь радиус D^2 равен квадрату расстояния от центра $\hat{\mu}$ до максимально удаленной точки, т. е. самая удаленная точка наблюдений находится на поверхности эллипсоида. Отметим, что в этой постановке объем существенно зависит от значения самой удаленной точки. Для оценки зависимости объема от расстояния приведем 9 самых удаленных наблюдений от выборочного центра $\hat{\mu}$ в порядке убывания квадрата расстояния: 38.9, 27.55, 19.35, 11.52, 11.42, 11.15, 7.96, 7.78, 6.9.

Объем эллипсоида вычисляется по известной формуле:

$$\frac{\pi^{m/2}}{\Gamma(m/2+1)} \frac{D^m}{\sqrt{\det(Q)}}, \quad \text{где } \Gamma(\cdot) \text{ — гамма-функция.} \qquad (15)$$

Если $D^2 = 38.9$, то объем эллипсоида равен 80% от объема параллелепипеда P_0. В то же время аналогичное отношение для построенного выше минимаксного параллелепипеда равно 7% (2-ой метод). Очевидно, что такая неэффективность эллипсоида является следствием множителя D^m в формуле его объема. Если отбросить самую удаленную точку, тогда $D^2 = 27.55$, и указанное отношение составляет уже 40%, а при удалении трех точек (тогда $D^2 = 11.52$) имеем отношение 6.8% и получаем меньший объем эллипсоида сравнительно с минимаксным параллелепипедом.

Заметим, что точки с большими квадратами расстояния от центра эллипсоида, равными 38.9 и 27.55, имеют в данном примере экономический смысл. Точка с расстоянием 38.9 — это 53-е наблюдение в выборке данных, соответствующее четвертому кварталу 2008 года, т. е. началу кризиса 2008 года, а точка с 27.55 — это четвертый квартал 1998 года, т. е. начало кризиса 1998 года.

В данном эксперименте существенное преимущество минимаксного параллелепипеда над эллипсоидом в задаче локализации многомерных данных, возможно, объясняется тем, что распределение исходных показателей значительно отличается от нормального (см. рис. 3). Для сравнения этих методов в случае *показателей, форма распределения которых приближается к нормальному закону,* были использованы данные из источника (Ферстер, Ренц, 1985). Наблюдения представляют выборку по 52 предприятиям из двух показателей:

x_1 — стоимость основных фондов;

x_2 — объем производства за квартал.

Объем параллелепипеда и эллипсоида оценивался аналогично эксперименту с 4-мя показателями, описанному выше. В результате объем параллелепипеда, построенного вторым минимаксным методом, оказался равным 27% от объема P_0, что меньше объема эллипсоида, равного 34%. На рисунке 4 приведено визуальное представление результатов эксперимента.

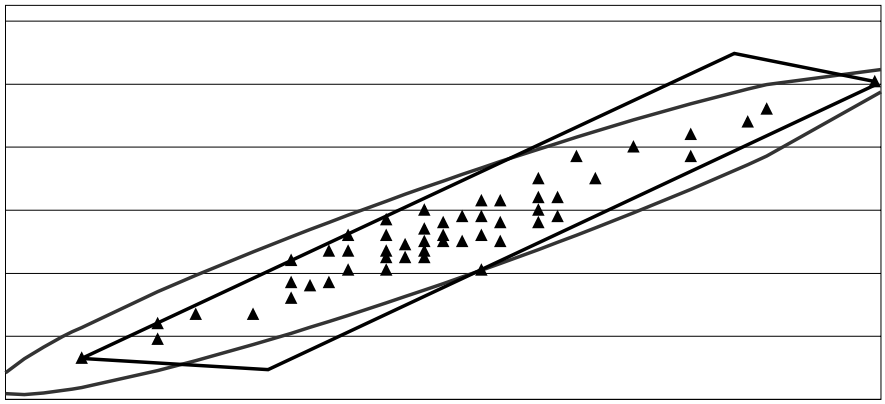

Рис. 4. Вид эллипсоида и параллелепипеда, включающих все наблюдения

4. Заключение

1. Рассмотренные методы вычисления главных компонент показали в численных экспериментах результаты (распределение изменчивости по главным компонентам и их направление), которые близки оценкам классического метода наименьших квадратов. В задаче локализации многомерных наблюдений эти методы дают лучшие результаты.

2. Вычислительная простота и наглядность метода максимального размаха делают полезным его применение на стадии предварительного анализа эконометрических данных, также он интересен как некоторый альтернативный взгляд на данные при использовании классического метода.

3. Второй минимаксный метод помимо определения главных компонент перспективно использовать в задачах локализации многомерных данных и оценки общей линейной связи показателей.

4. Теоретические свойства оценок в предложенных методах оценивания главных компонент пока не изучены, но, как известно, в одномерном случае минимаксная оценка (середина

выборочного размаха) при равномерном распределении случайной величины имеет скорость сходимости $1/n$, тогда как выборочное среднее значение в модели нормального распределения имеет скорость $1/\sqrt{n}$. Таким образом, можно ожидать эффективность рассмотренных методов в моделях данных с высокой степенью неопределенности, в частности, при равномерном законе распределения наблюдений внутри некоторого параллелепипеда.

Список литературы

Айвазян С. А., Бухштабер В. М., Енюков И. М., Мешалкин Л. Д. (1989). *Прикладная статистика. Классификация и снижение размерности*. М.: Финансы и статистика.

Гольштейн Е. Г. (1971). *Теория двойственности в математическом программировании и ее приложения.* М.: Наука.

Зиновьев А. Ю. (2000). *Визуализация многомерных данных*. Красноярск: Издательство Красноярского государственного технического университета**.**

Киселев Н. И. (1985). *Линейное программирование в экстремальных задачах статистики.* Ученые записки по статистике, т. 49. М.: Наука.

Ферстер Э., Ренц Б. (1983). *Методы корреляционного и регрессионного анализа.* М.: Финансы и статистика.

Asuncion A., Newman D. J. (2007). *UCI Machine Learning Repository.* http://archive.ics.uci.edu/ml/. Irvine, CA: University of California, School of Information and Computer Science.

Fisher R. A. (1936). The use of multiple measurements in taxonomic problems. *Annual Eugenics*, 7 (II), 179–188.